BIOINFORMATICS
Vol. 23 ISMB/ECCB 2007, pages i313–i318
doi:10.1093/bioinformatics/btm182

Biases induced by pooling samples in microarray experiments

Tristan Mary-Huard[1,*], Jean-Jacques Daudin[1], Michela Baccini[2,3], Annibale Biggeri[2,3] and Avner Bar-Hen[1]

[1] UMR AgroParisTech/INRA MIA 518, 16 rue Claude Bernard 75231 Paris Cedex 5 (France) and [2]Department of Statistics 'G. Parenti', viale Morgagni 59, 50100 and [3] CSPO Biostatistics Unit, via San Salvi, Florence, Italy

ABSTRACT

Motivation: If there is insufficient RNA from the tissues under investigation from one organism, then it is common practice to pool RNA. An important question is to determine whether pooling introduces biases, which can lead to inaccurate results. In this article, we describe two biases related to pooling, from a theoretical as well as a practical point of view.

Results: We model and quantify the respective parts of the pooling bias due to the log transform as well as the bias due to biological averaging of the samples. We also evaluate the impact of the bias on the statistical differential analysis of Affymetrix data.

Contact: maryhuar@inapg.fr

1 INTRODUCTION

In microarray experiments, pooling refers to the study design in which material collected from several individuals is combined in a pooled sample before hybridization. Labelling and hybridization are then performed on the composite sample. There are several reasons for pooling. When extraction from a single individual does not provide enough material, pooling is an alternative to RNA amplification (Gold *et al.*, 2004). Pooling is sometimes used to assemble a stable reference condition, to reduce the number of arrays for cost-saving purposes (Churchill, 2002), or to reduce the subject-to-subject variability and thus increase the power of statistical tests (Churchill, 2002; Churchill and Oliver, 2001; Han *et al.*, 2004; Simon and Dobbin, 2003).

Pooling design has recently received thoughtful attention in both statistical and biological publications about gene expression experiments. Authors mainly focus on two important questions:

- how to define equivalent designs i.e. what is the required number of subjects and arrays to achieve a given power in the statistical analysis (Shih *et al.*, 2004; Wit and McClure, 2004)?
- is the signal derived from a pool design equivalent to the average of expression signals from an individual-based design? This hypothesis, known as biological averaging assumption (BAA), has been studied on real data (Han *et al.*, 2004; Kendziorski *et al.*, 2003, 2005).

*To whom correspondence should be addressed.

In this article, we focus on the study of the validity of BAA. There is no general agreement in the literature. For example, in Kendziorski *et al.* (2005), the authors conclude that 'biological averaging occurs for most but not all genes'. However in Shih *et al.* (2004), the authors find that 'this assumption may not hold especially when the signals are strong...the pooling bias appears to be severer for the Affymetrix arrays'. However, no quantitative study of a possible pooling bias has yet been made.

There are two reasons why BAA may not hold (Kendziorski *et al.*, 2005):

- there may be an imperfect averaging of the individual RNA: X_p is different from $(1/n_s)\sum_{i=1,n_s} X_i$, where n_s is the number of samples, X_i is the number of labelled and hybridized RNA copies of a given gene from sample i for $i = 1$, n_s and X_p is the corresponding quantity for the pooled sample. We call this bias the *pool bias*.
- differences between the pool signal and the average of the individual signals could be due to the log transformation that occurs in the normalization process [for instance in the RMA procedure, Irizarry *et al.* (2003)]. Indeed, the log transformation is applied to individual samples in the absence of pooling, and to the pool sample otherwise, and, for any positive sequence $X_1,\ldots X_n$, $\frac{1}{n}\sum \log(X_i) \leq \log(\frac{1}{n}\sum X_i)$. If there is no *pool bias*, $X_p \simeq (1/n)\sum X_i$ and, as a result, the same equality cannot be true on the log scale, which is used for further statistical analysis. We call this bias the *log bias*.

The overall difference, on the log scale, between the pool and the mean of the corresponding individuals is called the *pooling bias*.

The goal of this study is to provide a better insight on the *pooling bias*, to quantify the respective parts of the *log bias* and the *pool bias* and to evaluate their impact on the statistical differential analysis of Affymetrix data. In Section 2, we define a general model for the expression measurement at the probe level. This framework is used in Section 3 to derive some tools to detect the two biases. We exemplify the two biases on both simulated and real data, using the Kendziorski experiment (Kendziorski *et al.*, 2005). Finally, Section 4 is devoted to the impact of the *pooling bias* on the differential analysis.

2 GENERAL FRAMEWORK

We consider here a two stage model. In the following the superscript B (respectively T) denotes the biological (respectively technical) variability.

2.1 Model on the expression of genes

The number of RNA copies of gene k from sample i is modelled by:

$$Y_{ik} = \lambda_k + \varepsilon^B_{ik}, \tag{1}$$

where λ_k is the population mean number of RNA copies of gene k, and ε^B_{ik} represents an independent random term with mean 0 and SD σ^B_k, corresponding to the subject-to-subject variability. σ^B_k is assumed to be finite, but may take different values for different genes. The number of labelled RNA copies of gene k of sample i, hybridized on the array i, (the numbering of the sample and of the array are identical) is modelled by:

$$X_{ik} = \alpha_i Y_{ik} = \alpha_i(\lambda_k + \varepsilon^B_{ik}), \tag{2}$$

where α_i is called the *efficiency factor*, and depends on the number of cells included in the RNA preparation and the quality of the hybridization and labelling processes. In the following, this *efficiency factor* is assumed to depend only on the sample RNA preparation and the array i, and does not depend on the gene and the probe. For a pool of n_s samples, the number of RNA copies of gene k contained in the pooled sample is

$$X_{pk} = \alpha_p(\lambda^{(p)}_k + \varepsilon^B_{pk}). \tag{3}$$

If BAA is true, then $\lambda^{(p)}_k = \lambda_k$ for all k.

2.2 Model on the measure of fluorescence at the probe level

For probe j associated with gene k, the expression measurement, on the log scale, is for the perfect match

$$\log(\mathrm{PM}_{ijk}) = \log(X_{ik}) + a_{jk} + \varepsilon^T_{ijk} \tag{4}$$

$$= \log\alpha_i + \log(Y_{ik}) + a_{jk} + \varepsilon^T_{ijk}, \tag{5}$$

where a_{jk} is the specific effect of probe j for gene k, and ε^T_{ijk} is an independent random term with mean 0 and SD σ^T, corresponding to the technical variability. The distribution of ε^T_{ijk} is assumed to be the same for each probe, each gene and each slide, and the two sources of variability ε^T_{ijk} and ε^B_{ik} are supposed independent.

At this step the model is quite general since few assumptions are made on ε^B_{ik} and ε^T_{ijk}. Notice that if we note $e_{ik} = \log(Y_{ik})$, the previous model can be rewritten

$$\log(\mathrm{PM}_{ijk}) = \log\alpha_i + e_{ik} + a_{jk} + \varepsilon^T_{ijk},$$

which is the model used in RMA normalization (Irizarry *et al.*, 2003). For a pooled sample we have:

$$\log(\mathrm{PM}_{pjk}) = \log\alpha_p + \log(Y_{pk}) + a_{jk} + \varepsilon^T_{pjk}. \tag{6}$$

3 BIAS QUANTIFICATION

3.1 Pool bias

We consider an experiment where RNA samples are extracted from n_s subjects. The RNA are used to make both individual samples and a pooled sample of the n_s subjects. Each sample is hybridized on a slide. For a given probe j and gene k, we obtain the measurements PM_{ijk} and PM_{pjk}, respectively for each individual i, $i = 1,\ldots, n_s$ and for the corresponding pool p. The mean expression for a given probe across the samples is:

$$\begin{aligned}\overline{\mathrm{PM}}_{jk} &= \frac{1}{n_s}\sum_{i=1}^{n_s}\mathrm{PM}_{ijk}\\ &= \frac{1}{n_s}\sum_{i=1}^{n_s}\alpha_i(\lambda_k + \varepsilon^B_{ik})e^{a_{jk}}e^{\varepsilon^T_{ijk}}.\end{aligned}$$

Since ε^B_{ik} and ε^T_{ijk} are independent, the mean of $\overline{\mathrm{PM}}$ with respect to the random variables ε^T_{ijk} and ε^B_{ik} is:

$$\begin{aligned}E_{B,T}\left[\overline{\mathrm{PM}}_{jk}\right] &= \frac{1}{n_s}\sum_{i=1}^{n_s}[\alpha_i E_B(\lambda_k + \varepsilon^B_{ik})]e^{a_{jk}}E_T(e^{\varepsilon^T_{ijk}})\\ &= \lambda_k e^{a_{jk}}\frac{1}{n_s}\sum_{i=1}^{n_s}\alpha_i E_T(e^{\varepsilon^T_{ijk}})\\ &= \lambda_k e^{a_{jk}}\alpha_\cdot E_T(e^{\varepsilon^T_{111}})\end{aligned}$$

where $\alpha_\cdot = \frac{1}{n_s}\sum_i \alpha_i$, since $\varepsilon^T_{ijk}, i = 1; ..., n_s$ are identically distributed. For a pooled sample,

$$\mathrm{PM}_{pk} = \alpha_p(\lambda^{(p)}_k + \frac{1}{n_s}\sum_{i=1}^{n_s}\varepsilon^B_{pk})e^{a_{jk}}e^{\varepsilon^T_{pjk}}.$$

The expectation of PM_{pk} is

$$\begin{aligned}E_{B,T}\left[\mathrm{PM}_{pjk}\right] &= \alpha_p E_B(\lambda^{(p)}_k + \frac{1}{n_s}\sum_{i=1}^{n_s}\varepsilon^B_{pk})e^{a_{jk}}E_T(e^{\varepsilon^T_{pjk}})\\ &= \alpha_p\lambda^{(p)}_k e^{a_{jk}} \times E_T(e^{\varepsilon^T_{111}})\end{aligned}$$

From these two computations we conclude that

$$r^{(p)}_{jk} = \frac{E_{B,T}\left[\overline{\mathrm{PM}}_{jk}\right]}{E_{B,T}\left[\mathrm{PM}_{pjk}\right]} = \frac{\lambda^{(p)}_k\alpha_p}{\lambda_k\alpha_\cdot} \tag{7}$$

If $\lambda^{(p)}_k = \lambda_k$ and if $\alpha_\cdot = \alpha_p$, *i.e.* the *efficiency factors* are similar for the pool and the individual slides, then $E_{B,T}[\overline{\mathrm{PM}}_{jk}]/E_{B,T}[\mathrm{PM}_{pjk}] = 1$.

Simulated data are an effective way to illustrate the theoretical computations above. Simulations are performed according to Models(2) and (5) with $\alpha_p = \alpha_i = 1$. For a given individual i and a given gene k, we have chosen to set $\lambda_k = 1000 + k$, $\sigma^b_k = \sqrt{k\sigma^B}$. Figure 1 a plots the pooled PM versus the mean of individual PMs. The values of the parameters are $n_s = 5$, $\sigma^2_T = 0.1$ and $\sigma^2_B = 0.1$ with normal distribution for the errors. The strong linearity along the line $y = x$ between PM_p and $\overline{\mathrm{PM}}$ illustrates the theoretical computations above. We observe similar results for different values of the parameters, provided that σ_T is not too high (figures not shown).

We now turn to real data, using the Kendziorski dataset (Kendziorski *et al.*, 2005). This experiment aims at comparing gene expression in mammary glands from female rats fed with two different diets (normal, denoted *A* and supplemented with the retinoic X receptor ligand LG100268, denoted *B*). RNA samples were obtained for 12 rats in each condition, and hybridized on Affymetrix RAE230A chips to measure gene expression for 15 923 genes. Individual RNA

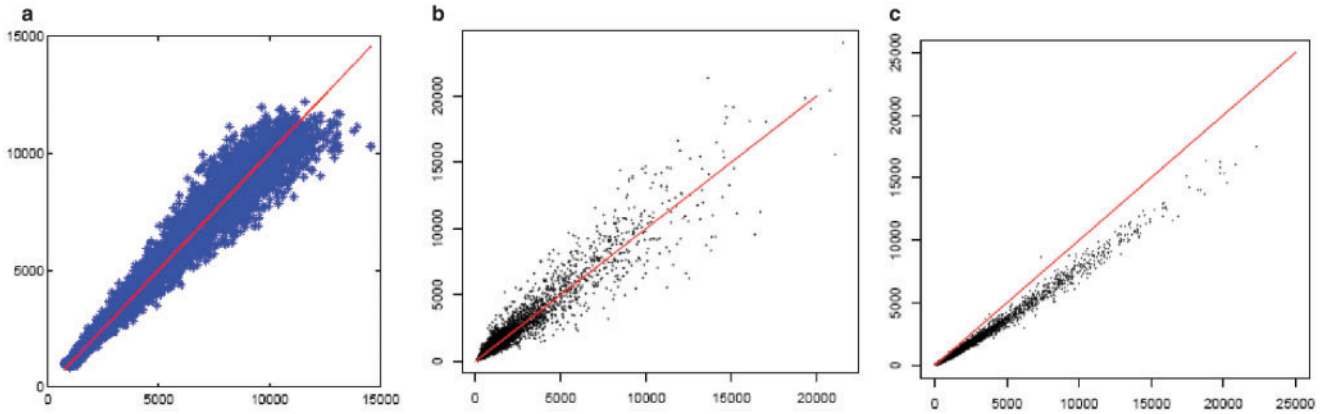

Fig. 1. Plot of the pool PM versus the mean of individual PM. Data are simulated according to Models (2) and (5) with gaussian errors. **(b)** PM of individual slide A2 versus PM of individual slide A3 (10 000 points presented in place of the total 175 477). **(c)** Plot of the pool of 12 PM versus the mean of individual PM (10 000 points presented).

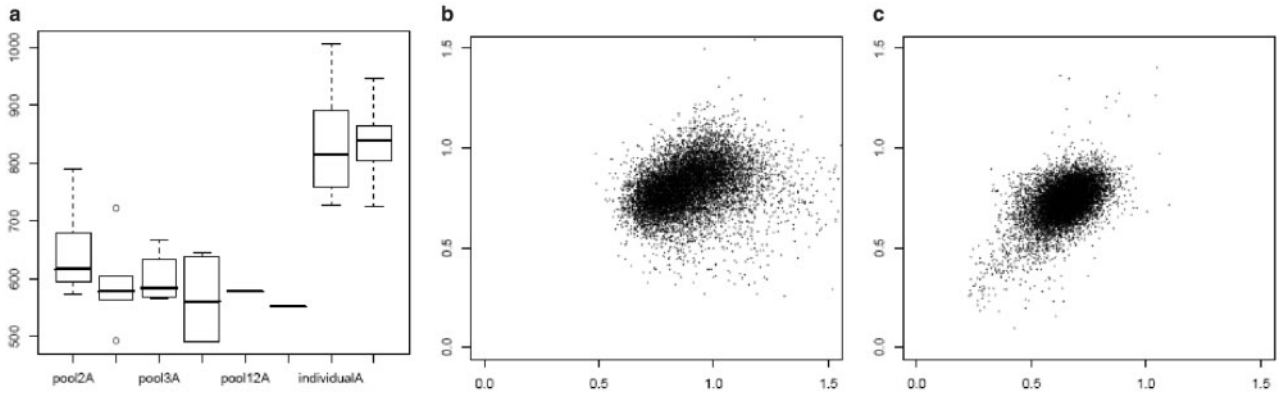

Fig. 2. (**a**) Boxplots for the mean PM of the arrays for pooled samples and individual samples (conditions A and B, before normalization). (**b**) plot of $r_{jk}^{(1)} = \mathrm{PM}_{pjk}^{(1)}/\mathrm{PM}_{\cdot jk}^{(1)}$ versus $r_{jk}^{(2)} = \mathrm{PM}_{pjk}^{(2)}/\mathrm{PM}_{\cdot jk}^{(2)}$, where the upper index refers to the specific pool of 3 considered (10 000 points presented in place of the total 175 477). (**c**) Plot of $r_{jk}^{(3)} = \mathrm{PM}_{pjk}^{(3)}/\mathrm{PM}_{\cdot jk}^{(3)}$ versus $r_{jk}^{(4)} = \mathrm{PM}_{pjk}^{(4)}/_{\cdot jk}^{(4)}$ where the upper index refers to the specific pool of 3 considered (10 000 points presented).

were also used to construct 6 pools of pairs, 4 pools of triples and 1 pool of 12 subjects, in each condition. Further details about this experiment can be found in Kendziorski *et al.* (2005).

If we compute the ratio between two individual slides 1 and 2, the same reasoning as above gives:

$$r_{jk}^{1,2} = \frac{E_{B,T}[\mathrm{PM}_{1jk}]}{E_{B,T}[\mathrm{PM}_{2jk}]} = \frac{\lambda_k \alpha_1}{\lambda_k \alpha_2} = \frac{\alpha_1}{\alpha_2}$$

This ratio only depends on the *efficiency factor* of the two slides, and should be roughly equal to 1. Figure 1b shows that the PM values of individual A2 versus A3 are distributed along the line $y = x$. The whole set of $11 \times 12 = 132$ individual ratios varies between0.75 and 1.38, the mean is 1.02 and the SE 0.13. This confirms that expression (7) gives a good picture of the biological process.

We observe a different picture for pool data, since the ratio $E_{B,T}[\overline{\mathrm{PM}}_{jk}]/E_{B,T}[\mathrm{PM}_{pjk}]$ is less than one. In Figure 1c, we consider the individual and pool arrays with $n_s = 12$. There appears to be a strong linear relationship, but the slope of the line is 0.75 rather than 1. This low ratio is not due to array effects, for similar computations on the 21 remaining pools of pairs, triples and 12 individuals give ratio values less than one. Except for one pool of pairs, the ratios are lower than 1 and lie between 0.5 and 0.96. The mean ratios are 0.77 for pools of 2, 0.715 for pools of 3 and 0.725 for pools of 12. These results are in keeping with the lower level of the mean expression for the arrays corresponding to pooled samples in comparison with arrays for individual samples (Fig. 2a).

One may wonder whether the ratio is probe (or gene) specific or not. Figure 2b and c represents the plot between ratios $r_{jk}^{(1)}$ and $r_{jk}^{(2)}$ for two different pool samples (pool of 3). Hence ratios vary greatly from pool to pool for a given probe. The Spearman correlations between these ratios from pool to pool are equal to 0.29 for the pools of 12 and for pools of 3 the Spearman correlations are between −0.34 and 0.43 with a mean equal to 0.14. For pools of 2, the mean correlation is equal to 0.08 and lie between −0.58 and 0.60. This argues for a non-specific ratio for a majority of probes.

We computed 2 scores per probe, S_{high} and S_{low}, that count how often a probe ratio $r_{jk}^{(p)}$ belongs to the 5% highest (respectively lowest) ratios for a given pool p. S_{low} and S_{high} vary between 0 and 22, since we considered the 22 pools all together. To get the same information at the gene level, we computed two scores S_{high}^g and S_{low}^g per gene, by summing the scores S_{high} and S_{low} of their associated probes. Each gene is represented by 11 probes (except for 36 that are represented by 20), meaning that S_{high}^g and S_{low}^g vary from 0 to 242. Under the assumption that no gene is specifically affected by pooling, each score has a binomial distribution $\mathcal{B}(n_0,p_0)$ with $n_0 = 22 \times 11$ and $p_0 = 0.05$. A gene will have a significant specific *pool bias* if S_{high} or S_{low} is greater than the 0.05/15 923 quantile of the binomial distribution, which is 30. In the Kendziorski data, we found more than 700 genes with either a S_{low}^g or a S_{low}^g value higher than 30, representing 4.6% of the total number of genes.

There are two reasons why the observed ratio $r_{jk}^{(p)}$ is not 1. Since

$$r_{jk}^{(p)} = \frac{\alpha_p}{\alpha_.} \times \frac{\lambda_k^{(p)}}{\lambda_k},$$

either $\alpha_p/\alpha_. \neq 1$ or $\lambda_k^{(p)}/\lambda_k \neq 1$. Of course it could be a combination of both reasons. From Figure 2 and the Spearman test we conclude that for a majority of probes the ratio $r_{jk}^{(p)}$ is similar, meaning that pooling induces a overall lowering effect that is not gene specific. This overall lowering corresponds to $\alpha_p/\alpha_. \neq 1$: there is a difference in efficiency between individual and pool slides. However, we showed that 4.6% of the genes were affected by an additional specific effect that comes from $\lambda_k^{(p)}/\lambda_k \neq 1$. While there is good hope that the normalization process eliminates the overall lowering effect of pooling (since most normalizations correct for a slide effect), the gene specific effect will not be removed and may affect the differential analysis.

3.2 Log bias

In this section, we study the bias that exists between the log-transformed pool signal and the mean of the log transformed individual signals. We define the *log bias* for a given probe j as:

$$B_{\log} = \log(\mathrm{PM}_{pj}) - \frac{1}{n_s}\sum_{i=1}^{n_s}\log(\mathrm{PM}_{ij})$$

According to (5), and dropping the gene index k for the sake of simplicity, we have for a given gene:

$$\begin{aligned} B_{\log} &= \left(\log(X_p) + a_j + \varepsilon_{pj}^T\right) - \frac{1}{n_s}\sum_{i=1}^{n_s}\left(\log(X_i) + a_j + \varepsilon_{ij}^T\right) \\ &= \log(X_p) + \varepsilon_{pj}^T - \frac{1}{n_s}\sum_{i=1}^{n_s}\log(X_i) - \frac{1}{n_s}\sum_{i=1}^{n_s}\varepsilon_{ij}^T. \end{aligned}$$

We suppose that the BAA hypothesis holds, and again for the sake of simplicity that $\alpha_. = \alpha_p$. We obtain:

$$B_{\log} = \log\left(\frac{1}{n_s}\sum_{i=1}^{n_s}X_i\right) + \varepsilon_{pj}^T - \frac{1}{n_s}\sum_{i=1}^{n_s}log(X_i) - \frac{1}{n_s}\sum_{i=1}^{n_s}\varepsilon_{ij}^T.$$

Assuming that the coefficient of variation of X_i is low, i.e. $\frac{\sigma_B}{\lambda} \ll 1$, and using $\log(1+t) \approx t - t^2/2$ we get:

$$\begin{aligned} B_{\log} &= \log\left[\lambda\left(1 + \frac{1}{n_s}\sum_{i=1}^{n_s}\frac{\varepsilon_i^B}{\lambda}\right)\right] - \frac{1}{n_s}\sum_{i=1}^{n_s}\log\left[\lambda\left(1+\frac{\varepsilon_i^B}{\lambda}\right)\right] \\ &\quad + \varepsilon_{pj}^T - \frac{1}{n_s}\sum_{i=1}^{n_s}\varepsilon_{ij}^T \\ &\approx \left[\frac{\sum_{i=1}^{n_s}\varepsilon_i^B}{\lambda n_s} - \frac{1}{2}\left(\frac{\sum_{i=1}^{n_s}\varepsilon_i^B}{\lambda n_s}\right)^2\right] - \frac{1}{n_s}\sum_{i=1}^{n_s}\left[\frac{\varepsilon_i^B}{\lambda} - \frac{(\varepsilon_i^B)^2}{2\lambda^2}\right] \\ &\quad + \varepsilon_{pj}^T - \frac{1}{n_s}\sum_{i=1}^{n_s}\varepsilon_{ij}^T \\ &\approx \frac{1}{2n_s}\sum_{i=1}^{n_s}\frac{(\varepsilon_i^B)^2}{\lambda^2} - \frac{1}{2}\left(\frac{\sum_{i=1}^{n_s}\varepsilon_i^B}{\lambda n_s}\right)^2 + \varepsilon_{pj}^T - \frac{1}{n_s}\sum_{i=1}^{n_s}\varepsilon_{ij}^T \\ &\approx \frac{1}{2\lambda^2}\left[\frac{1}{n_s}\sum_{i=1}^{n_s}(\varepsilon_i^B)^2 - \left(\frac{1}{n_s}\sum_{i=1}^{n_s}\varepsilon_i^B\right)^2\right] + \varepsilon_{pj}^T - \frac{1}{n_s}\sum_{i=1}^{n_s}\varepsilon_{ij}^T. \end{aligned}$$

Therefore

$$E_{BT}(B_{log}) \approx \frac{n_s - 1}{2n_s}\frac{\sigma_B^2}{\lambda^2}, \tag{8}$$

where n_s is the number of individuals combined in the pool sample. Thus, the expression is higher for the pool sample than for the mean of the corresponding individual samples on the log scale, and the mean difference, for a given gene, is proportional to the square of the coefficient of variation of $\epsilon_i^B, cv <= \sigma_B/\lambda$. Kendziorski found that for 25% of the genes with the largest SD, more than 80% have larger values in the pools of two (Fig. 4A in Kendziorski's paper). This artefact is well explained by the distortion of the log transformation that is described in this section.

Finally pooling results first in a lowering (for the Kendziorski experiment) of the raw microarray response that varies from microarray to microarray and secondly in an increase in the signal on the log scale after normalization, which depends on the biological variability of each gene. However, the sum of these two opposite effects is not equal to zero and varies on a relatively large scale from gene to gene and sample to sample. Therefore there is some theoretical and material evidence on the Kendziorski dataset that pooling affects the absolute measurement of gene expression.

4 DIFFERENTIAL ANALYSIS

An important point is to assess the consequences of pooling on the differential expression (DE) inference. First, if the arrays are made with pools composed of the same number of samples, the *efficiency factor* α_p is more or less the same for all arrays. Moreover, array to array normalization corrects, at least partially, the mean effect of the pooling on raw PM values, so that the *efficiency factor* should not distort the DE inference. However, the *log bias* artefact is not corrected by normalization precisely because it is produced by normalization. For example, if the mean and the variability of the expression of a particular gene are increased in condition A in reference with condition B, its DE will be higher in a pooled

experiment than in an individual one. Therefore individual experiments and experiments with pooling may lead to different conclusions for this given gene. One may think of other combinations of DE and variability that may lead to conflicting results.

To assess the impact of the *pooling bias* on DE, we performed a differential analysis on individual and pool arrays. For the individual study, RMA normalization was performed on the total batch of individual arrays, and genes were then ranked according to their associated T statistic, giving a unique reference list of genes. For the pool study, we performed three different normalization procedures for each pool batch (pool of 2, 3 and 12):

- Norm1 consists in a simple RMA normalization.
- Norm2 is a two-step normalization procedure. Data are first corrected for the *pool bias*: ratios $r_{jk}^{(p)}$ are estimated for each probe using the corresponding individual samples, and according to expression (7). Then classical RMA normalization is applied on the *pool bias* corrected data.
- Norm3 is a two-step normalization procedure, where data are first corrected for both the *pool bias* and the *log-bias* according to expressions (7) and (8), and then normalized with RMA.

For each of the three normalized datasets, we ranked the gene according to their T statistic to obtain three DE lists.

To compare our results with those of Kendziorski, we plotted the number of DE calls in common between the pool of 2 and reference for lists of fixed size (Fig. 3). We see that the correction of the *pool bias* increases the agreement between the reference and the pool of 12 lists. The additional correction of the *log bias* improves the agreement but the gain is very slight.

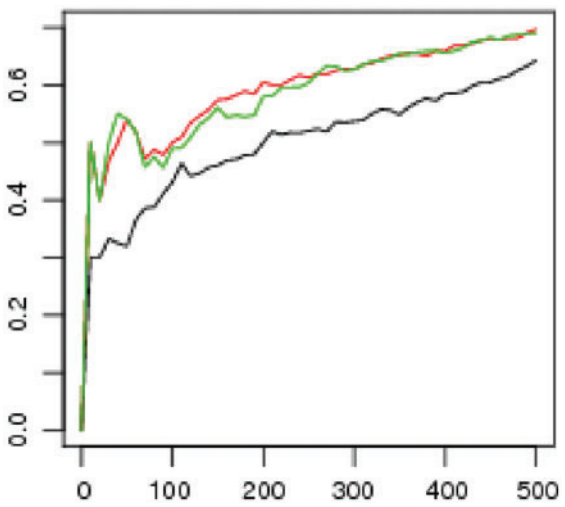

Fig. 3. Plot of the proportion of common DE genes between the individual analysis and the Norm1 normalized pool analysis (black curve), the Norm2 normalized pool analysis (red curve) and the Norm3 normalized corrected pool analysis (green curve), versus the size of the DE list. The pool analyses are performed on the total batch of pools of 2.

Results obtained with pools of 3 and pools of 12 are similar (not shown here).

5 DISCUSSION

In many articles, the BAA refers both to the fact that 'RNA abundance levels average out when pooled', and that 'average on the scale or raw RNA abundance will not correspond to the processed RNA measurement' (Kendziorski *et al.*, 2005). Here, we proposed to break down the overall *pooling bias* into two parts, the *pool bias* and the *log bias*, that are properly defined by expressions(7) and (8). This distinction allowed us to describe the *pool bias* as a combination of an overall effect which depends on the *efficiency factor* of each slide, and a gene specific effect which can be related to the RNA abundance with and without pooling. We were also able to quantify each part of the *pooling bias*. The main conclusions of this study are the following:

- pooling seems to lower the efficiency of the labelling or hybridization steps. This artefact, which has been found in Kendziorski's experiment, has to be confirmed by other experimental results. Shih *et al.* (2004) suggested that 'a possible reason for this artefact is that mixing of the RNA may cause some alteration of individual RNA contributions.' Such a bias can be easily detected in experiments by producing a few individual slides to compare their average signal level to that of pool slides. The impact of this bias on the differential analysis should be negligible, since it is mainly corrected by the array to array normalization step.
- some genes (up to 4.6% of the total number of genes in the Kendziorski experiment) are specifically affected by pooling. Specific gene biases are much more difficult to quantify and correct and would require both pool and individual slides for the same individuals, which cannot routinely be done in practice. Such biases could lead to different results between pool and individual slides analyses.
- the bias induced by the log transformation, included in most normalization methods, is experimentally and theoretically well assessed. While this bias is systematic, we showed that its consequences are of limited importance in the Kendziorski experiment. Yet it may potentially have an influence in other experiments, particulary for genes with high biological variability.

In Kendziorski *et al.* (2005), the authors observed that 'for the majority of genes where there was a large [*pooling bias*], the difference was similar across biological conditions'. Considering the two parts of the bias, it is difficult to conclude whether this may hold for further experiments. On one hand, for the bias to be 0 the specific gene bias has to be similar in both conditions. On the other hand, we showed that for a given gene, the *log bias* depends on the coefficient of variation cv. Compared with the log-ratio computed for individuals, the log-ratio for pools is biased by

$$cv_1^2 - cv_2^2 = \frac{\sigma_{B1}^2}{\lambda_1^2} - \frac{\sigma_{B2}^2}{\lambda_2^2}$$

For a non-DE gene, $\lambda_1 = \lambda_2$ and the bias is 0 if $\sigma^2_{B1} = \sigma^2_{B2}$. But for a DE gene $\lambda_1 > \lambda_2$, and the bias is 0 only if σ^2_{B1} increases in proportion to λ_1^2.

While the theoretical formulas we derived here are universal, the computations were all made on a unique set of microarrays. Future work will consist in applying the presented analytical tools to additional data to check to what extent the conclusions we have drawn on the Kendziorsky experiment can be generalized.

Conflict of Interest: none declared.

REFERENCES

Churchill,G.A. (2002) Fundamentals of experimental designs for cdna microarrays. *Nat. Genet.*, **32**, 490–495.

Churchill,G.A. and Oliver,B. (2001) Sex, flies and microarrays. *Nat. Genet.*, **29**, 322–356.

Gold,D. *et al.* (2004) A comparative analysis of data generated using two different target preparation methods for hybridization to high-density oligonucleotide microarrays. *BMC Genomics*, **5**.

Han,E.S. *et al.* (2004) Reproducibility, sources of variability, pooling, and sample size: important considerations for the design of high density oligonucleotide array experiments. *J. Gerontol. Biol. Sci.*, **59**, 306–315.

Irizarry,R.A. *et al.* (2003) Summaries of affymetrix genechip probe level data. *Nucleic Acids Res.*, **31**.

Kendziorski,C. *et al.* (2005) On the utility of pooling biological samples in microarray experiments. *Proc. Natl Acad. Sci. USA*, **102**, 4252–4257.

Kendziorski,C.M. *et al.* (2003) The efficiency of pooling mrna in microarray experiments. *Biostatistics*, **4**, 465–477.

Shih,JH. *et al.* (2004) Effects of pooling mrna in microarray class comparisons. *Bioinformatics*, **20**, 3318–3325.

Simon,R.M. and Dobbin,K. (2003) Experimental design of DNA microarray experiment. *BioTechniques*, **34**, S16–S21.

Wit,E. and McClure,J. (2004) *Statistics for Microarrays*. John Wiley & Sons, Hoboken, NJ, USA.